DOCUMENT RESUME

ED 163 935 IR 006 479

AUTHOR Regazzi, John J.
TITLE Indexing and Online Searching of Multi-Purpose Textual Databases: Conflict or Confluence?
PUB DATE Jun 78
NOTE 17p.; Paper presented at the Annual Meeting of the Special Libraries Association (Kansas City, Missouri, June 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Cost Effectiveness; *Data Bases; *Indexing; *Information Retrieval; Information Storage; Permuted Indexes; *Relevance (Information Retrieval); Search Strategies; Thesauri; Vocabulary
IDENTIFIERS Foundation Directory; Foundation Grants Index

ABSTRACT

The indexing and online searching of data bases have both conflicting and complementary structures, determined in part by the choice of indexing language. In the planning of a machine readable system, the Foundation Center, administrator of two data bases, initially chose a natural language for the lower cost, more rapid implementation, and greater efficiency than that of a controlled vocabulary. However, the input cost minimization afforded by the natural language alternative may result in cost shifting rather than reduction by increasing the cost of searching. Searchers also found the use of natural language undesirable due to the imprecision of the language, existence of many synonymous terms, and the difficulty of transfer of search strategies across systems. The decision was made to construct and initiate a controlled vocabulary index. The process of descriptor compilation and validation included rank ordering of present free-text descriptors' frequency, identification of synonyms, identification of relationships (e.g., broader term, narrower term), expert validation, classification of related trees, and construction of the thesaurus review process. (JVP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED163935

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

INDEXING AND ONLINE SEARCHING OF MULTI-PURPOSE TEXTUAL DATABASES: CONFLICT OR CONFLUENCE?

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John J. Regazzi

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

By: John J. Regazzi
Director, Documentation Services
The Foundation Center

Presented at: Special Libraries Association
Annual Conference, 1978

IR006479

INDEXING AND ONLINE SEARCHING OF MULTI-PURPOSE TEXTUAL DATABASES: CONFLICT OR CONFLUENCE?

Introduction

Arthur Koestler, in his recent book entitled Janus,[1] discusses the relationship between the nature of inquiry and hierarchies. Though Koestler was not specifically considering online searching or indexing in this context, there is an analogy which could be drawn among these concepts, and Koestler's discussion may provide a broad framework within which the relationship of indexing and searching may be considered. He states:

> "One obvious point is that hierarchies do not operate in a vacuum, but interact with others. This elementary fact has given rise to much confusion. If you look at a well-kept hedge surrounding a garden like a living wall, the rich foliage of the entwined branches may make you forget that the branches originate in separate bushes. The bushes are vertical, arborizing structures. The entwined branches form horizontal networks at numerous levels. Without the individual plants there would be no entwining, and no network. Without the network, each plant would be isolated, and there would be no hedge and no integration of functions."[2]

It is generally agreed that online searching and indexing are complementary functions. These functions, however, interact in a variety of ways and often serve different objectives, particularly in a multi-purpose information system. This paper presents a history of the development of two such multi-purpose textual databases--the Foundation Grants Index and the Foundation Directory. In tracing the development of these systems, emphasis will be given to the indexing procedures and vocabularies which are used and how the retrieval and dissemination of this information

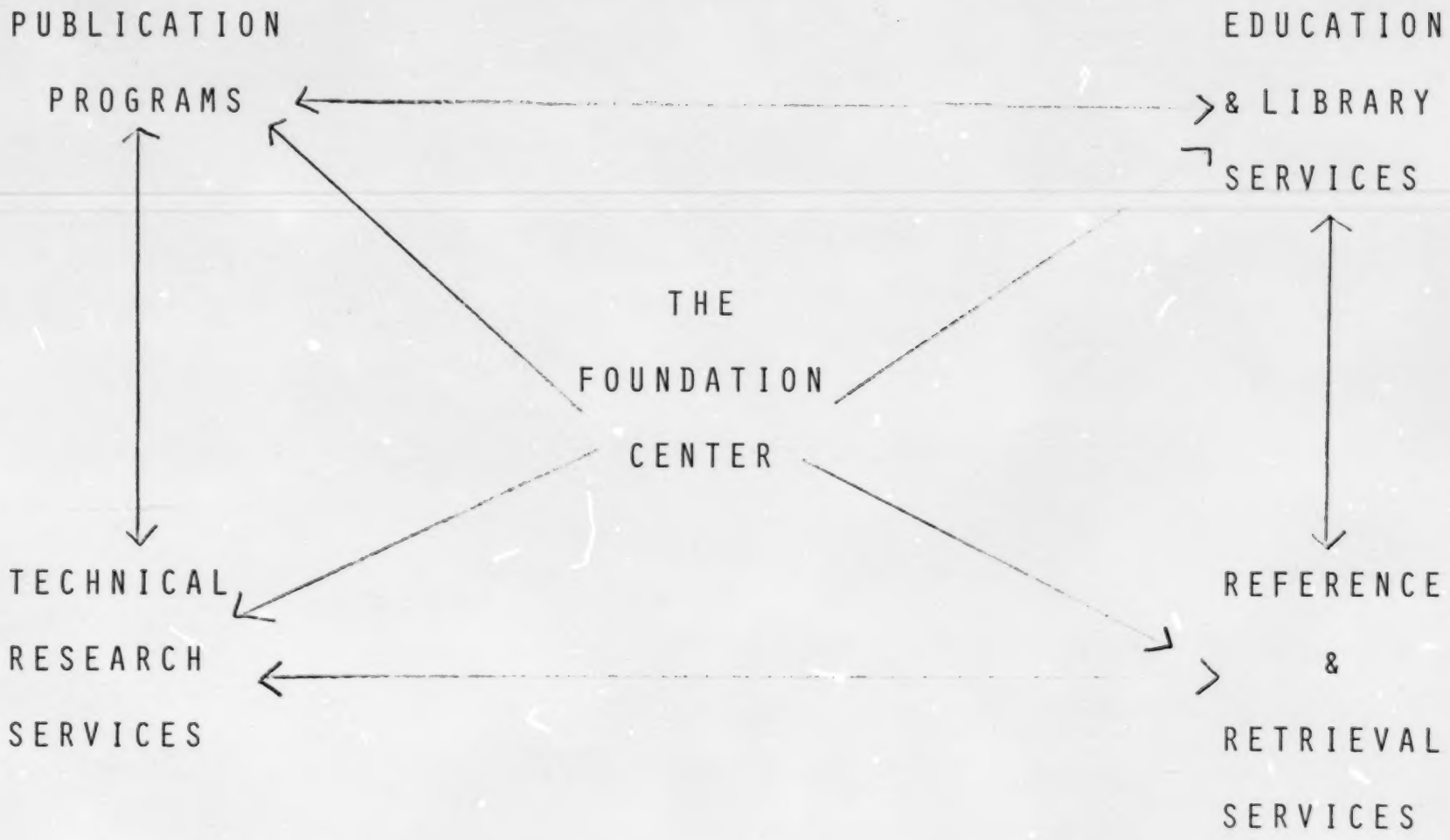

FIGURE 1. THE INTERRELATED FUNCTIONS OF THE FOUNDATION CENTER

has been affected by these indexing policies.

To this end, the paper is divided into four parts:

(1) the organizational setting in which the systems developed;

(2) a brief summary of development for both systems;

(3) a discussion of why a natural language for each system was chosen intially;

(4) and finally, an indication of why The Foundation Center is now considering the development of a controlled vocabulary.

The Organization: The Foundation Center

The Foundation Center, chartered as an educational corporation in 1956, is the only independent, not-for-profit organization in the United States dedicated entirely to the gathering, analysis, and dissemination of factual information on philanthropic foundations. Its two main purposes, of equal importance, are: (1) to be a useful informational resource for anyone interested in applying to grant-making foundations for funds; and (2) to compile reliable descriptive data and statistics on the foundation field for the use of foundation trustees and officers, regulatory agencies, and other interested organizations and individuals. Figure 1 represents a schematic of the four basic interrelating programs of The Foundation Center.

In order to meet these purposes, the Center has implemented a variety of free and fee-based services. The Center operates two libraries in New York and Washington, D.C. and has regional cooperating collections in 45 states, Mexico and Puerto Rico. These libraries are open to the public; have qualified staff; and provide services and information, including pre-prepared computer searches, at no cost to the patron. The

Center also maintains an Associates Program, for which there is a basic annual fee, in order to provide: (1) specialized information services by mail or telephone, (2) special compilations of data by library staff, and (3) custom computer searches by retrieval specialists. Foundations themselves may request any of these services for their own staff or board use by writing, telephoning, or visiting the Center or one of its branch offices located in Cleveland and San Francisco. Descriptive/statistical data is also supplied to foundations, governmental agencies, or research institutions on patterns of foundation funding, such as subject areas, types of support, recipient organization, and geographical distributions.

The Center has also developed a publications program to disseminate information to the wide spectrum of users concerned with foundation information. All the Center's major publications, including The Foundation Directory and The Foundation Grants Index, are computer generated from one or more machine-readable files.

Development of the Automated Information Systems

In 1972, The Foundation Center began to convert its manual files of foundation grants and entity information into machine-readable form. The conversion was undertaken for primarily two reasons: (1) to provide ready access to this information through the use of computer technology, and (2) to take advantage of the economics of automated photocomposition techniques for the publications mentioned above. By 1974, both the Grants Database and Foundation Directory Database were fully functioning. Figures 2 and 3 are schematic representations of the services and by-products from these systems and to what other systems each is related.

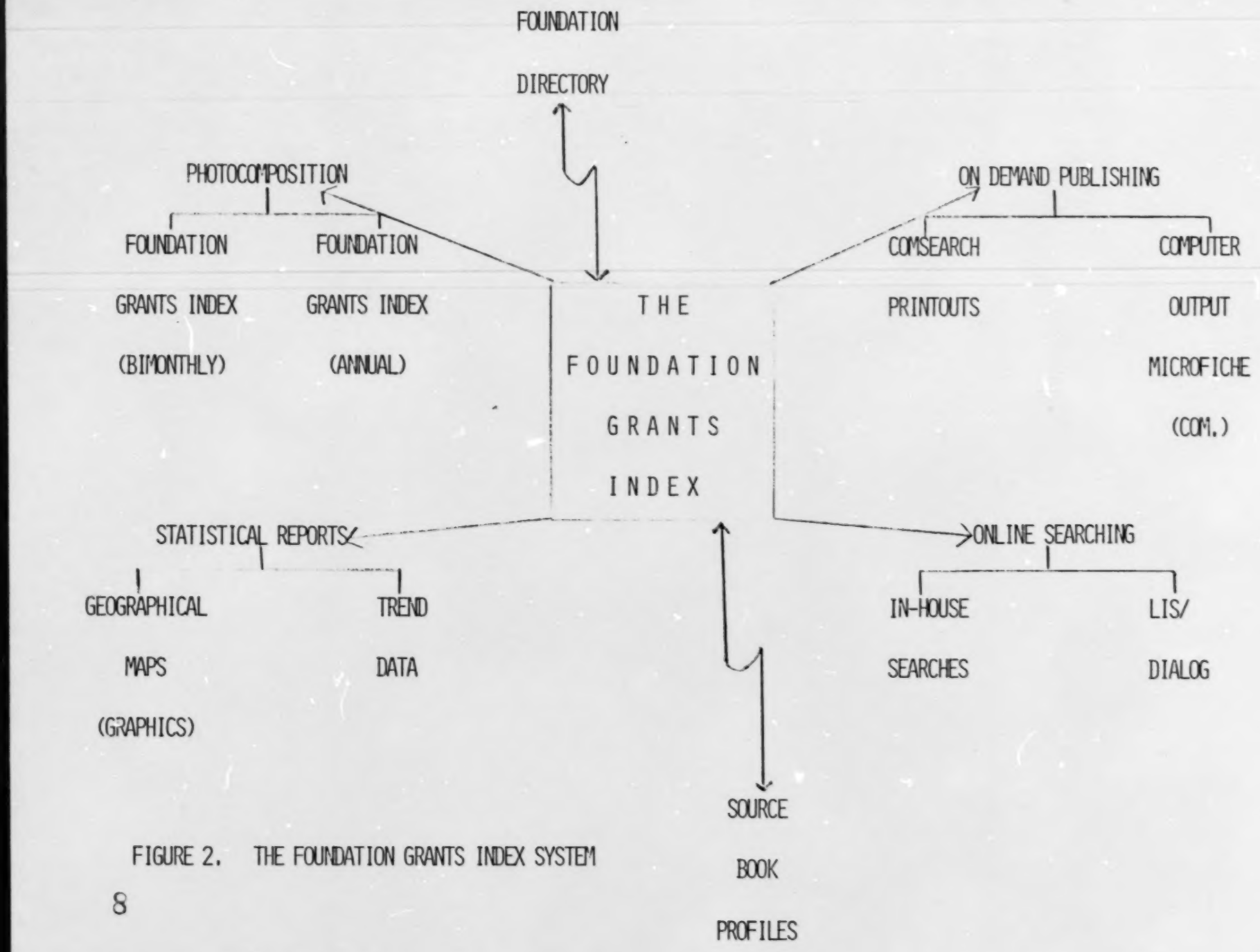

FIGURE 2. THE FOUNDATION GRANTS INDEX SYSTEM

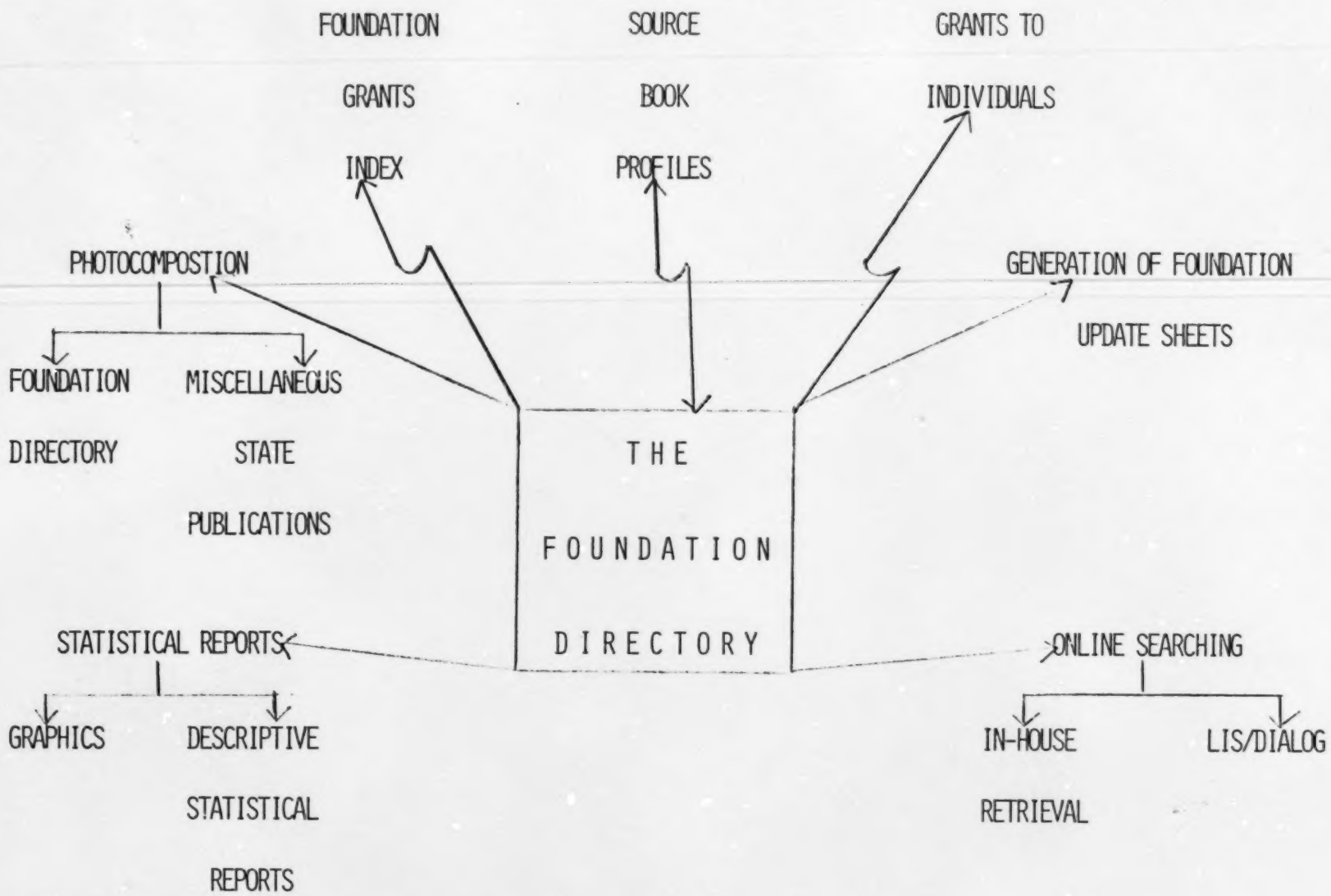

FIGURE 3. THE FOUNDATION DIRECTORY SYSTEM

Natural Language Indexing

During the intitial phase of implementation, the decision was made to rely upon essentially natural language descriptors and the use of a free vocabulary for the indexing and retrieval of data in both systems. Several factors were considered in the decision to apply free vocabularies to these databases: (1) the costs and length of time in developing a standard vocabulary with little or no raw data in many subject fields militated against the development of such a vocabulary; (2) there was considerable disagreement about the benefits and quality of a controlled vocabulary and a free vocabulary and whether the costs in developing such a controlled tool could be justified; and (3) it was felt that the indexing needs of the system could be reviewed after the system was implemented, and the question of controlled vocabulary could be looked at again at some later point. Though these three factors generally influenced the Center's initial decision to use a natural language vocabulary, there were a number of trade-offs implicit in this decision which were also considered.

Generally, there is little consensus on the preferability of natural language over controlled vocabularies. Yet, such early studies, such as the Cranfield II tests[2] and the Aberystwyth experiment,[3] seem to support natural language as a more effective and efficient indexing methodology. Other studies have found no statistically different results in recall and precision when comparing manual indexing from a formal thesaurus with machine indexing of title and abstract. Finally, many proponents of controlled vocabulary indexing who do not believe natural language to be as effective for retrieval, recognize its favorable comparison on economic grounds.

The difficulties of natural language fall into two categories, i.e., limitations due to (1) the imprecision of the language, and (2) the wide variety and forms of natural language expression. These two limitations are also reflected in the Center's experience with the use of natural language.

First, from the initial development of its databases, the Center wished to provide a subject index to each of its publications, and the computer alone was not seen as an effective device for providing such an index. It was decided that editors, relying upon the terms in the entry itself, would assign several keywords to a descriptor field, and these descriptors would be sorted and used for a subject index. Under such a procedure, the Grants Index and Directory systems have developed somewhat in isolation, and the present indexes bear little or no resemblance to one another. Other Center publications have used a similar procedure, and again, because of the diversity of the information contained in the publications, the indexes are quite distinct. Users, therefore, of the Center's publications and library services may find it difficult to transfer access points from one publication to another. Second, a commonly recognized limitation of natural language begins with the insufficiency of titles as the basis of keyword indexing. Though the information content of titles may be improving, titles are much more useful in science and technology than in the social sciences and humanities. Since the interests of foundations cover a wide range of activities, natural language indexing has resulted in different levels of precision for different classes of grants for both the database and the published volumes. A third disadvantage of the natural language relates to the "scattering" of terms on a subject. Scattering results from the lack

of control of synonyms, forms of entry, and generic-specific relationships. The scattering caused by variant forms of words (e.g., full time, fulltime) or various forms of an individual term (e.g., library, libraries, librarian, librarianship) are not as significant as those caused by synonyms and related terms, but both effect online and manual subject searching adversely. Fourth, besides scattering, natural language indexing has some disadvantages in retrieving precisely from the content of documents. In online retrieval, a searcher, when relying only upon natural language, cannot easily distinguish different meanings of the same word or homographs for example. Finally, it is difficult to develop statistical trends in granting when relying solely upon natural language. The ways of expressing a concept, theory, or problem may change much more quickly than the area of interest being studied. Much of what is now called "systems analysis" is essentially the same as "work-flow analysis" of several years ago. Therefore, it seems more efficient to classify grants prior to attempting a statistical analysis.

The Construction of a Controlled Vocabulary

In reconsidering the development of a thesaurus for these systems, the Center faced two, not unrelated, concerns: (1) the scope of the thesaurus, and (2) the costs required to develop a useful vocabulary for these files. First, private foundation funding spans virtually every aspect of human endeavor. Therefore, a thesaurus developed for indexing and retrieving philanthropic information would need to cover a wide range of topics. Second, can the benefits derived from this vocabulary justify the costs? The Center has recently developed a small prototype thesaurus of several hundred terms for these systems using a methodology which seems to be sound and efficient.

The strategy used has been to attempt to incorporate the advantages of natural language found in grant and entity information with other existing vocabularies which may be more specialized. This approach is based upon the premises: (1) natural language terms used in describing the grants themselves are the best predictors of terms which will be used to request and retrieve grant information, and (2) the frequency of natural language terms will suggest the scope and level of generality required for the controlled vocabulary.

The methodology for developing this thesaurus is divided into six steps. First, a listing in rank order of words and bound terms is produced and reviewed. Second, from these terms sets of synonyms are identified, and the most appropriate or the most frequently occurring descriptor is chosen as the preferred terms. Other terms are used for lead-in descriptors. Third, the identification of the relationships (e.g., broader term, narrower term, related terms, use, and used for) of each term on the basic list are made and assigned. Fourth, it is realized that certain terms may not be included in this basic list. Other available thesauri and qualified persons are consulted, and additional or other preferred terms as well as their relationships are integrated into the basic list. Fifth, the main facets or "trees" under which the terms will be classified are identified and assigned. Sixth, a procedure for reviewing, integrating new words, and deleting old or little used terms from this thesaurus is established.

Several references have been made to possible benefits expected from the development of a thesaurus for these systems. To recapitulate, there are three major benefits of the thesaurus for The Foundation Center. First,

for certain types of online and manual searches, a controlled vocabulary will be more efficient, since the searcher is not required to assemble all synonyms and related terms, i.e., the searcher can use the relationships and terms already provided in a thesaurus. Second, users of the published volumes of these systems will have access to a common vocabulary rather than two substantially different sets of terms. This standardization, if uniformly applied to Center services, should provide easier subject access to those services. Third, trend data analysis could be developed for foundation funding. Though to a certain extent this is possible with a natural language system, it is a tedious task, and impossible to automate fully. With a thesaurus and clearly defined subject terms, trend data can be cumulated and consistently analyzed over a period of time.

Conclusion

The use of a natural language or a controlled vocabulary will probably vary from system to system. As previously discussed, both offer attractive and appropriate benefits. These benefits, however, must be reviewed in relation to the costs of providing them. Measuring the specific benefits of a controlled vocabulary is no easy matter. In economic terms, however, the natural language alternative provides an input cost minimization, and this procedure is generally justified on that basis. A question, however, is whether such a procedure is an actual cost reduction or a subtle cost shifting which in turn maximizes the costs of another function of the system, e.g., the output function of searching. If this is so where the user is willing and able to accept the incremental increase of output costs, the

system may function in an acceptable way. Where this is not the case, the quality and utility of the system may be compromised.

Finally, as Koestler has suggested for inquiry and hierachy, the functions of searching and indexing are complementary principles in the architecture of online information retrieval systems. The index must provide the tree structure, or aborizing function of the system; it is through this indexing structure that inquiries can engage, and entwine in various ways with the system. A controlled vocabulary seems to provide a structure which, if kept current, will enable the user to retrieve information in a creative and flexible way.